融合兴趣的微博用户相似度计算研究 *

黄贤英，阳安志，刘小洋，刘广峰

(重庆理工大学 计算机科学与工程学院，重庆 400054)

摘 要：针对传统基于用户的博文内容和共同好友数在计算微博用户的相似度时存在潜在误差过大的问题，而基于用户多源背景信息的相似度计算模型，有计算复杂度高且忽略了用户的兴趣等问题，提出了一种结合用户兴趣和背景信息的综合相似度计算方法(BIBS)。首先从用户的标签中提取用户的兴趣，当用户的标签缺失时，通过对用户关注关系网络中的重要用户聚类来间接获取用户的兴趣点，以此计算用户的兴趣相似度；其次根据用户的性别、年龄和地点等背景属性计算用户的背景相似度，层次化的挖掘出最相似的用户；最后基于新浪微博的数据进行实验分析。结果表明，与基于多源信息相似度的微博用户推荐算法(MISUR)相比，该方法在用时更少的情况下，准确率、召回率和 F 值分别提高了 8.1%、16.7%和 13.6%，证明了提出的 BIBS 方法的有效性和准确性。

关键词：微博；兴趣；用户聚类；相似度计算

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2018.07.0469

Research on similarity computation of microblog users combining user interests

Huang Xianying, Yang Anzhi, Liu Xiaoyang, Liu Guangfeng

(*School of Computer Science & Engineering*, *Chongqing University of Technology*, *Chongqing* 400054, *China*)

Abstract: The traditional method of calculating the similarity of the Microblog users based on the user's blog content and the number of common friends has the problem of excessive potential error, and the similarity calculation model based on the user's multi-source background information has high computational complexity and ignore the user's interest and other issues, the author puts forward a combined with user's interest and background information to calculate the comprehensive similarity (BIBS) . The method extracts the user's interest from the user's tag. When the user's tag is missing, the user's interest is indirectly obtained by clustering the important user in the user's attention network, and the user's interest similarity is calculated, and then the background similarity of the user is calculated according to the background information such as the gender, age and location of the user, so that the most similar users are hierarchically mined. Experiments and analysis based on the data of Sina Microblog show that compared with MISUR algorithm based on the similarity of multi-source information, the algorithm can improve the accuracy, recall rate and F-measure by 8.1%, 16.7% and 13.6% respectively with less time consuming, which proves the effectiveness and accuracy of the proposed BIBS method.

Key words: Microblog; interest; user clustering; similarity calculation

0 引言

随着信息技术的进一步提高，在线社交网络得到快速的发展，参与社交网络的用户也越来越多。据 CNNIC 发布的第 41 次《中国互联网络发展状况统计报告》[1]显示，截止 2017 年 12 月，微博用户超过 3.1 亿，年增长率达到 16.4%。愈加庞大的用户基数使得用户在微博中搜索信息、建立互动关系时，会因信息过载的问题而困惑。如何帮助用户在大量的人群节点中发现其兴趣点，这对于社交网络平台和用户都具有极其重要的意义，解决这个问题有效的方法之一就是个性化推荐。传统推荐领域的方法包括协同过滤推荐方法、基于内容的推荐方法和混合推荐方法等[2-3]，这些方法在好友推荐、新闻推荐、音乐推荐等方面有很多实际的应用。个性化推荐中一个很重要的研究是相似度计算方法[4]，如用户相似度计算、物品相似度计算等，它是为用户进行相关推荐的基础。大多数传统的推荐算法是根据用户对项目的历史评分数据，建立相应的用户兴趣模型，依此计算

收稿日期：2018-07-25；**修回日期：**2018-09-05 **基金项目：**重庆市教育委员会人文社会科学研究项目（17SKG144、18SKGH110）；国家教育部人文社科青年基金资助项目（16YJC860010）；国家社科基金资助项目（17XXW004）；2018 年重庆市科委技术创新与应用示范项目（cstc2018jscx-msybX0049）

作者简介：黄贤英（1967-），女，重庆人，教授，硕导，主要研究方向为计算机应用等；阳安志（1993-），男（通信作者），四川人，硕士研究生，主要研究方向为在线社交网络、机器学习（pureyangcry@foxmail.com）；刘小洋（1980-），男，安徽人，副教授，硕导，博士（后），主要研究方向为社交网络、信息传播与计算机应用等；刘广峰（1995-），男，山东人，硕士研究生，主要研究方向为社交网络，机器学习等.

用户的相似度，产生推荐结果。随着 Web 2.0 的快速发展，国外的 Twitter、Facebook，国内的新浪微博等在线社交网络的流行，促使传统推荐系统融合微博用户的背景信息和社会行为信息为用户进行相关推荐。

近年来，在微博推荐领域，提出了很多新的用户相似度计算方法，如徐志明等人[5]针对微博用户信息的特点，综合考虑用户的背景信息、微博文本和社交信息等属性来计算用户的相似度。文献[6,7]结合用户的性别、年龄及博文内容等信息，提出了基于余弦距离的用户相似度综合计算方法，而姚彬修等人[8]结合用户的博文内容、交互信息和共同粉丝数，提出了基于多源信息相似度的微博用户推荐算法。这些方法都综合考虑了用户的多方面信息来构建对应的特征向量，利用余弦距离来挖掘相似用户。但由于微博的博文内容有最大长度的限制，直接构建用户特征向量，利用余弦相似性不足以衡量微博用户的相似性[9]，此外还会有潜在误差过大、计算复杂度高等问题，而 He 等人[10]根据博文的转发关系网络对用户进行聚类，发现同一社区的用户有相似的兴趣，表明在社交网络中，用户的交际圈更多是建立在共同的兴趣上，结合用户的兴趣，能准确的发现社区中的相似用户。文献[11-13]都是基于用户的兴趣来计算用户的相似度。黄宏程等人[11]研究了微博用户的长、短时兴趣，利用兴趣相似度来预测用户的关系；陈杰等人[12]提出一种基于用户动态兴趣的社交网络的微博推荐方法。结合用户兴趣进行相关推荐变得越来越流行，Xing 等人[14]深入研究了用户自身的多方面信息，提出可以利用用户的博文内容、自定义标签及关注关系来挖掘用户的兴趣，表明微博用户的自定义标签比博文内容能更加准确的反映用户的实际兴趣。马慧芳等人[15]也深入研究了用户的自定义标签来为用户进行推荐。虽然基于标签的推荐更加准确、有效，但微博中有大量普通用户并没有自定义标签，文献[14,15]列举了提取用户兴趣的一些方法，如从用户个人资料和博文内容中提取兴趣，而仲兆满等人[16]研究发现，通过用户的关注关系间接获取用户兴趣的方法是合理、有效的，用户因对某个明星感兴趣，才会关注他，这体现了用户对该明星所在领域感兴趣。

本文分析了微博用户的关注关系网络结构，因为大量普通用户缺少代表其兴趣的自定义标签，所以提出利用用户关注关系中的重要用户来间接获取用户的兴趣的方法。首先利用 PageRank 算法挖掘出被关注的重要用户，然后对其进行聚类，间接的获取用户兴趣，最后构建了基于兴趣和背景信息的用户相似度计算方法 BIBS(calculation of similarity based on user's interest and background information)。实验结果表明，该方法能更加准确的计算微博用户的相似度。

1 相关研究

1.1 传统计算方法

在传统电子商务服务中，个性化推荐技术通过研究用户的兴趣爱好，为客户推荐其感兴趣的商品等资源。如基于用户的协同过滤推荐算法，根据用户对项目的评分矩阵，计算用户间的相似度，找出与目标推荐用户最近邻的用户集合，然后对最近邻居集合进行加权，最后产生目标用户的推荐集，此类算法能够有效地使用相似用户的反馈信息来为用户产生推荐结果[17]。随着社交网络的快速发展，个性化推荐技术也在社交网络中得到了不同程度的应用，微博领域中的相关推荐方法也越来越多。较早提出的方法是根据用户之间的共同邻居数量来计算微博用户的相似度，如共同邻居 CN（common neighbors）模型、Jaccard 相似度计算模型，CN 相似度模型的计算公式如下：

$$Sim(A,B)=\frac{|CN(A)\cap CN(B)|}{|CN(A)\cup CN(B)|} \tag{1}$$

其中：$Sim(A,B)$ 代表用户 A 、B 的相似度，$CN(A)$ 代表用户 A 的好友集合，$CN(B)$ 代表用户 B 的好友集合，用户 A 、B 的共同好友数越多，表明 A 、B 越相似。但这类算法推荐结果的准确性较差，一方面，它忽略了来自微博用户自身的信息，如用户的喜好、年龄等信息；另一方面，与现实朋友关系不同，社交网络中的用户不可能与好友列表中的每个用户有较强的联系，基于共同好友数产生的推荐结果，用户满意度较低。

1.2 融合多源信息方法

针对传统推荐算法存在的问题，研究人员开始结合微博用户自身的背景信息来计算用户的相似度，文献[5]考察了用户的背景信息、微博文本信息和社交信息来计算用户相似度，文献[6]提出结合用户背景信息和互动信息构成的综合相似度计算模型，而文献[8]首先将用户的博文内容进行预处理、分词，为了获得微博内容的关键词表，使用了一种用于信息检索与数据挖掘的加权技术 TF-IDF（term-frequency–inverse-document-frequency），利用余弦距离计算博文的内容相似度，再根据两个用户间对彼此微博的兴趣度来计算用户交互行为的相似度，最后基于用户双方的共同关注好友数和粉丝数来计算用户的社交关系相似度，提出了基于用户多源信息的相似度计算方法 MISUR（user recommendation algorithm based on the similarity of multi-source information），各部分的计算公式定义如式(2)~(5)。

$$sim_1(u,v)=\cos\left(m(u),m(v)\right)=\frac{\sum_{i=1}^{n}w_{ui}w_{vi}}{\sqrt{\sum_{i=1}^{n}w_{ui}^2}\sqrt{\sum_{i=1}^{n}w_{vi}^2}} \tag{2}$$

其中：$sim_1(u,v)$ 表示用户 u 、v 的博文内容相似度，$m(u)$ 、$m(v)$ 表示用户 u 、v 的博文文本向量。

$$sim_2(u,v)=\frac{\sum_{r\in blog}(u_r-\bar{u})(v_r-\bar{v})}{\sqrt{\sum_{r\in blog}(u_r-\bar{u})^2}\sqrt{(v_r-\bar{v})^2}} \tag{3}$$

其中：$sim_2(u,v)$ 表示用户 u 、v 的交互行为相似度，u_r 和 v_r 分别表示用户 u 、v 对共同交互过的微博 r 的兴趣度，$\bar{u}$ 和 $\bar{v}$ 表示用户 u 、v 对所有交互过的微博的兴趣的平均值。

$$sim_3(u,v)=w_1\times sim(Following(u),Following(v))+ w_2\times sim(Follower(u),Follower(v)) \tag{4}$$

$$w_1+w_2=1 \tag{5}$$

其中: $sim_3(u,v)$ 表示用户 u 、v 的社交关系相似度, $Following(u)$ 、$Following(v)$ 表示用户 u 、v 共同关注的好友数相似，$Follower(u)$ 、$Follower(v)$ 表示共同粉丝数相似， w_1 和 w_2 表示各部分的权重。

最后将用户的微博内容相似度、交互相似度和社交关系相似度三部分进行综合来计算微博用户的多源信息相似度。

这些计算方法都综合考虑了用户的多方面信息，利用余弦距离来计算微博用户的相似度，但仍有一些问题。如，用户的所发博文随机性较大，计算相似度会有潜在误差过大的问题，另外根据用户背景信息构建特征向量，直接利用 Cosine 来计算用户的相似度，在实际应用中的占用资源较多，计算复杂性相对较大。

1.3 用户兴趣挖掘

社交网络中用户的交际更多的是建立在共同兴趣之上，为了挖掘出微博用户的兴趣，Xing 等人[14]深入研究用户的标签信息，发现经过新浪微博官方认证的重要用户（加 V 用户）明显比普通用户倾向于为自己添加更多的标签，实验表明从标签信息中获取微博用户的兴趣的方法最有效；文献[11]研究了用户的长、短时兴趣，表明标签可以代表用户的长期兴趣，是相对稳定的。

虽然通过标签信息能准确的挖掘出用户的兴趣，很多明星大 V 也愿意为自己添加更多的标签，但普通用户很少为自己定义标签，所以只利用标签信息来挖掘微博用户的兴趣存在局限性。

研究表明，通过用户的关注关系间接获取用户兴趣的方法是有效的[16]，如果两个用户同时关注了明星谢娜（自定义标签有：社会闲杂人等、主持人），说明这两个用户可能都对主持人感兴趣，进一步可以预测出他们可能对娱乐节目、综艺等领域感兴趣。

因此，在用户自定义标签较少的情况下，相比于从用户博文内容中挖掘兴趣的方法，通过用户关注关系挖掘出用户兴趣更加准确，所以本文提出一种基于用户关注关系挖掘用户兴趣的方法，并依此来计算微博用户的兴趣相似度。

2 基于用户关注关系的挖掘兴趣

2.1 用户关注网络

微博社交网络中，如果用户对某个用户感兴趣，他可以关注此用户，也可以关注很多感兴趣的其他用户，同样的，其他用户也可以关注他，许多用户的相互关注就构成了关注关系网络。在这个关系网络中，一些用户节点因其自身的特点，被很多其他用户节点关注，实际中，这些用户通常是网络社区中的活跃分子，他们对其他用户节点有较大的影响力，这些重要用户被称为“意见领袖”(opinion leader)[18]。相比于一般用户，在用户关注的好友中，更能代表用户兴趣点的通常是网络中的这些重要用户。因为用户的关注关系网络较为复杂，为了挖掘出能代表用户兴趣的重要用户，采用 PageRank 页面排序算法。

2.2 重要用户挖掘

用户在社交网络中的关注关系可以被视为有向链接，著名的基于链接的排序算法之一是 PageRank 页面排序算法。该算法是由 Google 的两位创始人提出的，最初是为了实现网页排名，在搜索引擎中被广泛使用。页面的分数通过不断的迭代计算得到，但当某些页面只存在入链或出链时，迭代结果会出现“排名泄漏”和“排名下沉”的问题,得到不合理的排名结果。为了解决这个问题，引入了随机浏览模型，即每个页面都可以随机访问其他页面。算法最终的表示如下：

$$PageRank(x) = (1-d) + d\sum_{y \in L(x)} \frac{PageRank(y)}{N_y} \tag{6}$$

其中: $PageRank(x)$ 、$PageRank(y)$ 表示页面 x 和 y 的排名分数，$L(x)$ 表示 x 的链入页面集合， N_y 表示页面 y 总的链出数， d 是阻尼系数，表示一个页面被其他页面随机访问的概率，即使页面 x 没有被其他页面引用，也能获得的 $(1-d)$ 基本分数，保证页面的迭代分数能收敛。在挖掘重要用户时，公式中的 $PageRank(x)$ 、$PageRank(y)$ 表示用户 x 和 y 的重要度，$L(x)$ 可以表示用户 x 所关注的用户集合或互动关系集合等， N_y 表示对应集合的好友数。结合微博领域中用户的特点，大量的研究在挖掘重要用户时，都改进了 PageRank 算法，如曹玖新等人[18]通过改进 PageRank 算法来挖掘意见领袖。

本文利用 PageRank 算法在用户的关注网络中挖掘出最能代表该用户兴趣点的重要用户，间接获取用户的兴趣。

2.3 用户兴趣挖掘

首先通过 PageRank 算法挖掘出关注网络中的重要用户，提取重要用户的标签，构建重要用户的标签向量，以此对重要用户进行聚类得到聚类结果。聚类结果的类别向量定义为

$$(Cluster1, Cluster2, Cluster3, \cdots\cdots, ClusterN) \tag{7}$$

其中， $Cluster1$ 到 $ClusterN$ 表示不同的聚类类别。

然后统计用户在不同类别中关注的好友数，构建该用户的兴趣向量，定义如下：

$$Intertest(A) = (count1, count2, count3, \cdots\cdots, countN) \tag{8}$$

其中: $Intertest(A)$ 表示用户 A 的兴趣向量， $count1$ 表示用户 A 在类别 1 中关注的用户数，$countN$ 表示用户 A 在类别 N 中关注的用户数。但是当用户在某一类别中会有较多的关注用户，会干扰余弦相似度的结果。基于 TF-IDF 的思想，对其进行归一化，得到用户 A 的兴趣向量，如式(9)所示。

$$Intertest(A) = \left(\frac{count1}{Num_{c1}}, \frac{count2}{Num_{c2}}, \frac{count3}{Num_{c3}}, \cdots\cdots, \frac{countN}{Num_{cN}}\right) \tag{9}$$

其中: $Intertest(A)$ 表示用户 A 的兴趣向量， Num_{c1} 表示类别 1 中的所有用户数， Num_{cN} 表示类别 N 的中的所有用户数。

最后，利用余弦距离来衡量不同用户的兴趣相似度，计算公式如式(10)所示。

$$Sim_{Interest}(A,B)=\cos(Intertest(A),Intertest(B))=\frac{Intertest(A)\cdot Intertest(B)}{|Intertest(A)||Intertest(B)|} \quad (10)$$

其中：$Sim_{Interest}(A,B)$ 表示用户 A 和 B 的兴趣相似度，$Intertest(A)$ 、$Intertest(B)$ 表示 A 、 B 的兴趣向量。

为了进一步说明该方法的具体计算过程，举例如图 1 所示。

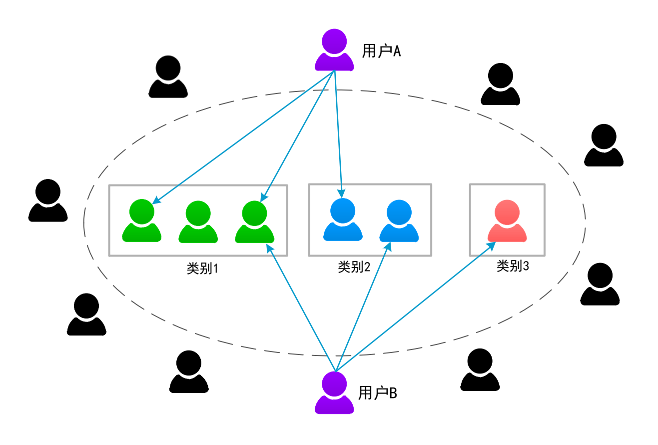


图 1　计算例图

Fig.1　Example of calculation method

如图 1 所示，将用户关注关系中的重要用户进行聚类，结果得到类别 1、2 和 3，其中用户 A 关注了类别 1 和 2 中的用户，用户 B 关注了类别 1、2 和 3 中的用户，因此用户 A 的兴趣向量为 $(2,1,0)$，用户 B 的兴趣向量为 $(1,1,1)$。然后对其进行归一化，得到用户 A 的兴趣向量为 $(2/3,1/2,0)$，用户 B 的兴趣向量为 $(1/3,1/2,1)$，最后通过余弦距离计算得到用户 A 、B 的兴趣相似度。

通过用户关注关系中的重要用户来间接获取用户的兴趣点，一方面可以在普通用户标签缺失较多情况下，挖掘用户的兴趣，进而为用户进行相关推荐；另一方面，与传统的直接利用用户的背景信息特征向量来计算用户相似度的算法相比，该方法不必对每一个用户进行计算，故所需的时间相对较少，复杂度明显降低。

3　综合相似度计算方法

3.1　兴趣相似度

与生活中朋友关系的建立不同，在社交网络中，用户因共同兴趣组成不同的社区，同一个社区中的用户通常会有相似的兴趣[10]。传统推荐领域中，研究人员从用户对商品、音乐的评分中挖掘用户的兴趣，如用户对某本书籍进行了评分，系统会找出与该书相似的书籍，推荐给用户。随着微博愈加热门，其资源和数据的进一步扩大，挖掘用户兴趣来为用户进行推荐的研究工作愈加重要，而利用微博用户关系来间接挖掘用户兴趣的方法较少。本文通过用户社交关系网络中关注的重要用户来挖掘用户的兴趣，从而计算用户的兴趣相似度。该方法具体过程见第 3 章的说明。

3.2　背景信息相似度

在计算微博用户的相似度时，很多研究都综合考虑了用户多方面的背景信息。文献[5]针对微博用户本身的信息，通过利用用户的位置信息、标签信息和个人描述信息来计算用户的背景相似度。在此基础上，文献[6]结合用户的性别、年龄和地理信息来计算用户的背景相似度。基于已有的研究，本文结合了用户的性别、年龄和地点信息来计算微博用户的背景相似度。

1)性别

性别往往是衡量一个人的重要标准，在微博领域中，不同性别用户的行为差别较大。如男性用户一般会对体育、科技、时政等方面的内容更感兴趣，而女性用户更可能会更关注美妆、综艺娱乐、减肥等方面的信息。用户性别属性的定义公式如下：

$$U_{sex}(A)=\begin{cases}1, & A=\text{“男”}\\ 0, & A=\text{“女”}\end{cases} \quad (11)$$

其中，$U_{sex}(A)$ 表示用户 A 的性别。

2)年龄

在社交网络中，不同年龄的用户往往差别较大。不同年龄层的用户往往拥有不同的经历、阅历和关注点，因此他们的相似度较小。一般而言，年龄差越小，年龄差占年龄的比例越低，用户的兴趣越接近，其相似度越高[6]。用户年龄属性定义如式(12)所示：

$$U_{age}(A)=\frac{age_A-age_{min}}{age_{max}-age_{min}} \quad (12)$$

其中:$U_{age}(A)$ 表示用户 A 的计算年龄，age_A 表示 A 的实际年龄，$age_{\max}$ 表示数据中的最大年龄值，$age_{\min}$ 表示数据中的最小年龄值。

3)地点

在实际应用的社交推荐系统中，有很多基于地点信息的推荐，如，附近的人的推荐。文献[19-20]都是基于地点信息来挖掘相似用户，并取得了不错的效果，结合地点信息的推荐受到越来越多的用户喜爱。微博中，用户的地点信息包括省份，地市等。用户的地点属性定义如下：

$$U_{address}=(u_{province},u_{city}) \quad (13)$$

其中:$U_{address}$ 表示用户的地点特征信息，$u_{province}$ 表示用户所在的省份，u_{city} 表示用户所在的城市，计算时需要转换成对应的数值。

综上所述，在分析用户多方面的背景信息后，结合用户的性别、年龄和地点信息构建用户背景信息向量，定义如下：

$$BI_A=(U_{sex},U_{age},U_{address}) \quad (14)$$

其中:BI_A 表示用户 A 的背景特征向量，U_{sex} 、U_{age} 、$U_{address}$ 分别表示用户 A 的性别、年龄、地点特征信息。

背景相似度计算公式为

$$Sim_{BI}(A,B)=cos(BI(A),BI(B))=\frac{BI(A)\cdot BI(B)}{|BI(A)||BI(B)|} \quad (15)$$

其中: $Sim_{BI}(A,B)$ 表示用户 A 和 B 的背景相似度，$BI(A)$ 、$BI(B)$ 表示用户 A 、B 的背景特征向量，通过余弦距离来计算用户的背景相似度。

3.3　综合相似度计算

系统分析了微博用户的个人资料、关注关系、互动关系以

及兴趣点，基于已有的研究，提出了结合用户兴趣相似与背景相似的综合模型来挖掘微博社交网络中的相似用户，模型如图 2 所示。

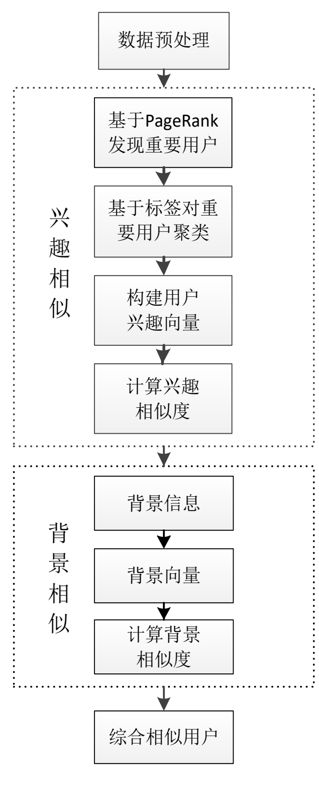


图 2 模型结构图

Fig.2 Structure diagram of the model

用户综合相似度的计算步骤：

a) 数据预处理；

b) 计算兴趣相似，挖掘兴趣最相似的 N 个用户；

c) 计算背景相似，从兴趣相似的 N 个用户中挖掘背景相似用户；

d) 综合相似的用户。

其中，N 的取值与用户规模数有关，不同的取值会影响计算结果的准确率。如，为某用户推荐 10 个最相似用户，若 N 的取值过小，无法保证能找到综合相似的前 10 个用户，但若 N 取值过大，会增加算法的计算复杂度，所以应根据实际情况，合理的选择 N 的取值。

因此在计算用户相似度时，先挖掘出兴趣相似的用户，再计算其背景相似度，层次化的挖掘出综合信息最相似的用户。该方法一方面降低了对所有用户计算特征向量的复杂性，提高了算法的性能；另一方面保证了推荐结果与用户兴趣点是相关的。

4 实验与分析

4.1 数据获取

为了验证提出模型在计算微博用户相似度的有效性，本文利用 UCI 官网的 MicroblogPCU 数据集(https://archive.ics.uci.edu/ml/machine-learning-databases/00323/)来进行实验。该数据集包括 59191 名用户以及 142369 条的关注关系信息，其中 782 名用户有详细个人信息，包括用户 ID、用户名、性别、账号等级、地点信息、标签、博文数、关注用户数、粉丝数，262 位用户带有自定义标签，标签总数为 1441 个，实验利用他们的关注关系来构建用户的关系网，剩下的用户作为测试来验证模型。

4.2 结果与分析

首先分析了用户的关注关系网络，得到用户的关注兴趣图，如图 3 所示。

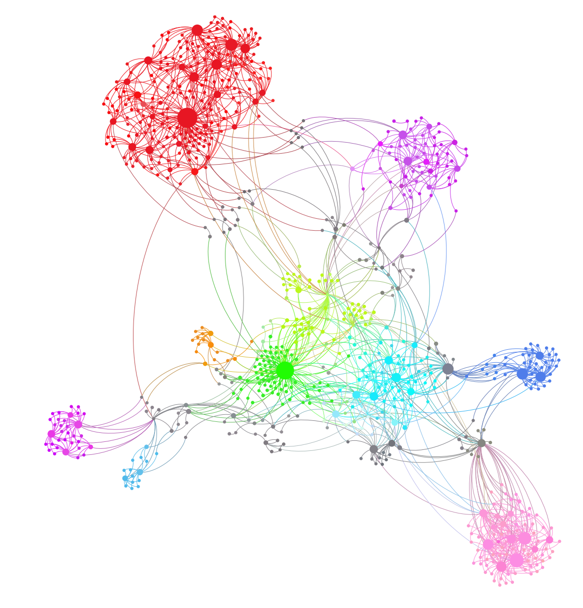


图 3 用户关注兴趣图

Fig.3 Interest diagram of user's attention

图 3 采用 ForceAtlas 布局来可视化用户的关注兴趣。图中的每个节点表示一个用户，每两个用户的关注形成一条边，被越多的用户关注，图中该用户的节点就越大。可以发现，该网络中用户的兴趣有明显的类别，大多数用户的兴趣相对集中；用户的兴趣有多个类别，一些用户同时关注了几个领域，其兴趣相对较广。对于关注多个领域的用户，传统 CN 算法在进行相关推荐时，会有潜在误差过大的风险，因为用户的兴趣是多样的，仅通过共同好友数量来推荐相似用户，效果较差。

首先通过 PageRank 算法，挖掘出该网络中被关注较多的重要用户，将这些用户的标签进行分词，构建标签向量，基于标签进行聚类。应该将重要用户聚成多少个类别？聚类结果的有效性评价标准有两种，一种是通过测量聚类结果和参考标准的一致性来评价聚类结果的优良；另一种是评价同一聚类算法在不同聚类数条件下，聚类结果的优良程度[21]。这里采用 Calinski-Harabasz(CH)指标来评价聚类结果的好坏。CH 指标通过类内离差矩阵的紧密度和类间离差矩阵的分离度来判断聚类结果的好坏，公式的定义如下：

$$CH(k)=\frac{trB(k)/(k-1)}{trW(k)/(n-k)} \tag{16}$$

其中：n 表示聚类结果数，k 表示当前的类，$trB(k)$ 表示类间离差矩阵的迹，$trW(k)$ 表示类内离差矩阵的迹。CH 值越大，同类的元素越紧密，不同类别越分散，聚类效果就越好。

实验中，不同聚类数的结果下 CH 值如图 4 所示。

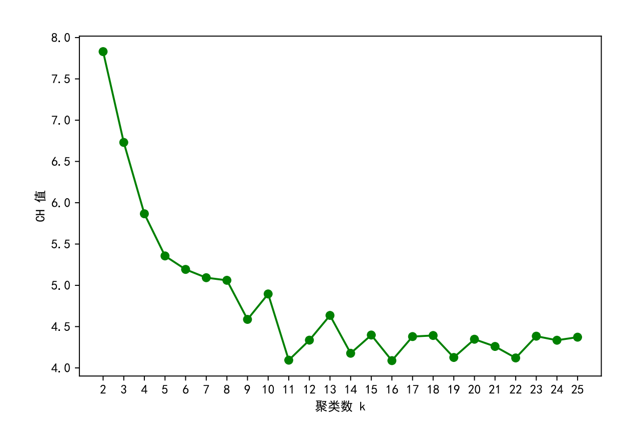


图 4　不同聚类数的 CH 值

Fig.4　CH of different cluster numbers

图 4 是将这些重要用户聚成 2 到 25 个不同类别的结果。可以看到，当 $k=5$ 时，CH 值较好。但发现依此建立用户的兴趣向量，较多被关注的用户在同一个类中，模型的准确率较差，当 $k=10$ 时，算法的准确率相对较好，因此将重要用户聚成 10 个类，并依此构建普通用户的兴趣向量，故不同数据集的聚类数要根据实际情况而定。

为了验证算法的有效性，采用准确率(precision rate)、召回率（recall rate）、*F* 值（*F*-measure）作为评估指标，各个公式定义如式(17)~(19)所示。

$$Precision = \frac{N_r \cap N_u}{N_r} \tag{17}$$

其中：N_r 表示向用户推荐的好友集合，N_u 表示用户已经关注的好友集合，*Precision* 表示准确率，是指向用户推荐的正确相似用户数与推荐用户数的比值。

$$Recall = \frac{N_r \cap N_u}{N_u} \tag{18}$$

其中：*Recall* 表示召回率，指向用户推荐的正确相似用户数与用户已经关注的好友数的比值。

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{19}$$

其中：$F-measure$ 表示正确率和召回率的调和平均值，$F-measure$ 值越大，则该方法的结果越准确。

对比算法包括综合用户兴趣相似与背景相似的计算方法(BIBS 算法)、只结合用户兴趣的相似度计算方法(BIS 算法)、文献[8]提出的 MISUR 算法以及共同邻居数算法(CN 算法)。实验从数据集中选取了 131 名用户及其关注好友关系数据，验证了本文的算法和其他对比算法的准确率、召回率和 F 值。

各个算法的准确率对比结果如图 5 所示。

如图 5 所示，推荐人数从 5 到 25 名用户，综合用户兴趣相似与背景相似的计算方法(BIBS 算法)的准确率是最高的，说明从用户关系网络中获取用户兴趣的方法是有效的，综合用户兴趣和背景信息能更加准确的挖掘出微博中的相似用户，相比于 MISUR 算法，准确率平均提高了 8.1%。

各个算法的召回率对比结果如图 6 所示。

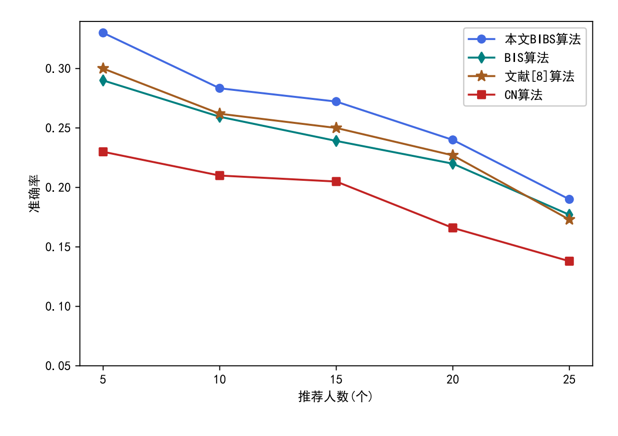


图 5　准确率对比图

Fig . 5　Comparison of precision rate

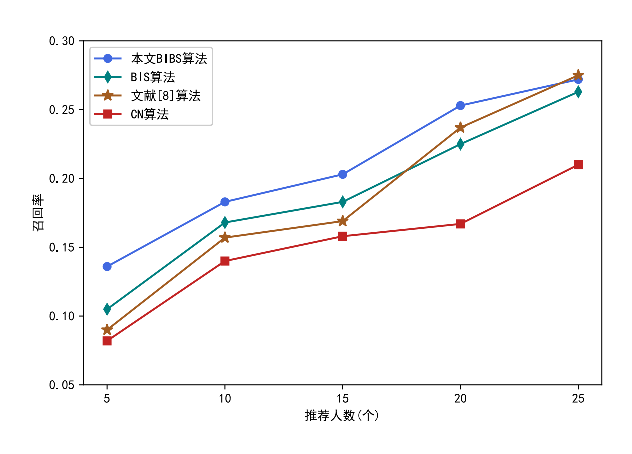


图 6　召回率对比图

Fig.6　Comparison of recall rate

从图 6 可以看出，随着推荐人数的增多，所有算法的召回率值都在上升，推荐人数为 25 时，文献[8]算法的召回率最高，但是本文的 BIBS 算法综合结果相对较好，召回率平均提高了 16.7%。

各个算法的 *F* 值对比结果如图 7 所示。

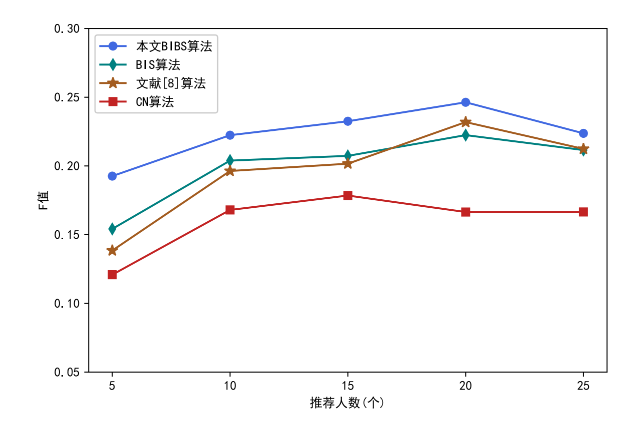


图 7　*F* 值对比图

Fig.7　Comparison of *F*-measure

F 值作为综合评价指标，根据图 7 所示，BIBS 算法是最有效的，相比于 MISUR 算法，F 值平均提高了 13.6%。综上所述，本文提出的基于用户兴趣和背景信息的综合用户相似度方法是优于其他对比算法的。

在对 131 名用户进行好友推荐时，统计了各个算法的运行时间，如图 8 所示。

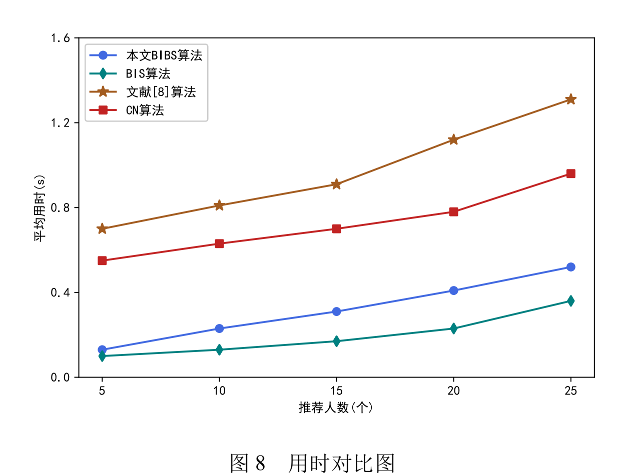


图 8　用时对比图

Fig.8　Comparison of time consumption

从图 8 中可以看出，相比文献[8]的 MISUR 算法和共同邻居 CN 算法，本文算法的用时最少。因为 MISUR 算法从用户的博文中提取特征，建立用户特征向量，再根据用户的背景信息，构建用户的背景特征向量，利用余弦距离来衡量用户的兴趣相似度，这样直接计算所有用户的相似度，用时相对较长，共同邻居 CN 算法同样如此，需要对所有用户进行共同好友数比较。而本文的算法，先从用户的关注关系中提取出用户兴趣，找到兴趣最相似的用户，再从兴趣相似的用户中计算背景相似度，层次化的找出最相似的用户来进行相关推荐，用时明显减少，具有良好的实用性。

5　结束语

已有的研究在计算微博用户的相似度时，综合考虑了用户的多源信息，建立用户背景信息特征向量，利用余弦距离来计算用户的相似度，这些方法存在潜在误差过大、计算复杂等问题，也忽略了用户的兴趣点，于此，本文提出基于用户兴趣与背景信息的综合用户相似度计算方法。文章首先概述了微博用户相似度计算的相关研究，接着介绍了一种基于用户关注关系间接获取用户兴趣点的方法。该方法利用 PageRank 算法从用户关注关系挖掘出最能代表用户兴趣点的重要用户，然后对这些重要用户进行聚类，间接获取用户的兴趣点。再结合用户的背景信息，根据用户的性别、年龄和地点信息来计算用户的背景相似度，最后提出了基于用户兴趣和背景信息的综合计算模型，层次化的挖掘出微博中的相似用户。在新浪微博的数据集上验证了该模型的有效性，准确率和性能得到显著提升，而用时较少。下一步工作是对算法进行优化和改进，从用户的转发、评论关系网络中挖掘出用户的短期兴趣，进一步提高算法在计算微博用户相似度的准确性。

参考文献：

[1] 中国互联网络发展状况统计报告 [R]，北京：中国互联网络信息中心，2018. (Statistical report on the development of China's Internet [R] , Beijing: China Internet Network Information Center, 2018.)

[2] 孟祥武，刘树栋，张玉洁，等. 社会化推荐系统研究 [J]. 软件学报，2015, 26 (6): 1356-1372. (Meng Xiangwu, Liu Shudong, Zhang Yujie, *et al.* Research on social recommender systems [J]. Journal of Software, 2015, 26 (6): 1356-1372.)

[3] 黄震华，张佳雯，田春岐，等. 基于排序学习的推荐算法研究综述 [J]. 软件学报，2016, 27 (3): 691-713. (Huang Zhenhua, Zhang Jiawen, Tian Chunqi, *et al.* Survey on learning-to-rank based recommendation algorithms [J]. Journal of Software, 2016, 27 (3): 691-713.)

[4] Wu Xiaokun, Huang Yongfeng. SigRA: a new similarity computation method in recommendation system [C]// Proc of International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. 2017: 148-154.)

[5] 徐志明，李栋，刘挺，等. 微博用户的相似性度量及其应用 [J]. 计算机学报，2014, 37 (1): 207-218. (Xu Zhiming, Li Dong, Liu Ting, *et al.* Measuring similarity between microblog users and its application [J]. Chinese Journal of Computers, 2014, 37 (1): 207-218.)

[6] 郑志蕴，贾春园，王振飞，等. 基于微博的用户相似度计算研究 [J]. 计算机科学，2017, 44 (2): 262-266. (Zheng Zhiyun, Jia Chunyuan, Wang Zhengfei, et al, Computing Research of user similarity based on micro-blog [J]. Computer Science, 2017. 44 (2): 262-266.)

[7] 段旭磊，张仰森，孙祎卓. 微博文本的句向量表示及相似度计算方法研究 [J]. 计算机工程，2017, 43 (5): 143-148. (Duan Xulei, Zhang Yangsen, Sun Yizhuo. Research on sentence vector representation and similarity calculation method about microblog texts [J]. Computer Engineering, 2017, 43 (5): 143-148.)

[8] 姚彬修，倪建成，于苹苹，等. 基于多源信息相似度的微博用户推荐算法 [J]. 计算机应用，2017, 37 (5): 1382-1386. (Yao Binxiu, Ni Jiancheng, Yu Pingping, *et al.* Micro blog user recommendation algorithm based on similarity of multi-source information [J]. Journal of Computer Applications, 2017, 37 (5): 1382-1386.)

[9] Pandey N. Density based clustering for Cricket World Cup Tweets using cosine similarity and time parameter [C]// Proc of India Conference. IEEE, 2016: 1-6.

[10] He Yuan, Wang Cheng, Jiang Changjun. Mining coherent topics with pre-learned interest knowledge in Twitter [J]. IEEE Access, 2017, 5 (99): 10515-10525.

[11] 黄宏程，陆卫金，胡敏，等. 用户兴趣相似性度量的关系预测算法 [J]. 计算机科学与探索，2017, 11 (7): 1068-1079. (Huang Hongcheng, Lu Weijin, Hu Min, *et al.* User Relationships prediction algorithm with interest similarity measurement [J]. Journal of Frontiers of Computer Science &

Technology, 2017, 11 (7): 1068-1079.)

[12] 陈杰，刘学军，李斌，等. 一种基于用户动态兴趣和社交网络的微博推荐方法 [J]. 电子学报, 2017, 45 (4): 898-905. (Chen Jie, Liu Xuejun, Li Bin, *et al.* Personalized microblogging recommendation based on dynamic interests and social networking of users [J]. Acta Electronica Sinica, 2017, 45 (4): 898-905.)

[13] Jain A, Gupta A, Sharma N, *et al.* Mining application on analyzing users'interests from Twitter [C]// Proc of International Conference on Internet of Things and Connected Technologies. 2018: 1-8.

[14] 邢千里，刘列，刘奕群，等. 微博中用户标签的研究 [J]. 软件学报, 2015, 26 (7): 1626-1637. (Xing Qianli, Liu Lie, Liu Yiqun, *et al.* Study on user tags in Weibo [J]. Journal of Software, 2015, 26 (7): 1626−1637.)

[15] 马慧芳，贾美惠子，张迪，等. 融合标签关联关系与用户社交关系的微博推荐方法 [J]. 电子学报, 2017, 45 (1): 112-118. (Ma Huifang, Jia Meihuizi, Zhang Di, *et al.* Microblog recommendation based on tag correlation and user social relation [J]. Acta Electronica Sinica, 2017, 45 (1): 112-118.)

[16] 仲兆满，管燕，胡云，等. 基于背景和内容的微博用户兴趣挖掘 [J]. 软件学报, 2017, 28 (2): 278-291. (Zhong ZhaoMan, Guan Yan, Hu Yun, *et al.* Mining user interests on microblog based on profile and content [J]. Journal of Software, 2017, 28 (2): 278-291.)

[17] 荣辉桂，火生旭，胡春华，等. 基于用户相似度的协同过滤推荐算法 [J]. 通信学报, 2014 (2): 16-24. (Rong Huigui, Huo Shengxu, Hu Chunhua, *et al.* User similarity-based collaborative filtering recommendation algorithm [J]. Journal on Communications, 2014 (2): 16-24.)

[18] 曹玖新，陈高君，吴江林，等. 基于多维特征分析的社交网络意见领袖挖掘 [J]. 电子学报, 2016, 44 (4): 898-905. (Cao Jiuxin, Chen Gaojun, Wu Jianglin, *et al.* Multi-feature based opinion leader mining in social networks [J]. Acta Electronica Sinica, 2016, 44 (4): 898-905.)

[19] Zou Zhiqiang, Xie Xingyu, Chao Sha. Mining user behavior and similarity in location-based social networks [C]// Proc of International Symposium on Parallel Architectures. 2016: 167-171.

[20] 丁勇，刘菁，蒋翠清，等. LBSN 中考虑用户交友偏好的好友推荐方法研究 [J]. 系统工程理论与实践, 2017, 37 (11): 2975-2982. (Ding Yong, Liu Jing, Jiang Cuiqing, *et al.* A study of friends recommendation algorithm considering users'preference of making friends in the LBSN [J]. Systems Engineering-Theory & Practice, 2017, 37 (11): 2975-2982.)

[21] 周开乐，杨善林，丁帅，等. 聚类有效性研究综述 [J]. 系统工程理论与实践, 2014, 34 (9): 2417-2431. (Zhou Kaile, Yang Shanglin, Ding Shuai, *et al.* On cluster validation [J]. Systems Engineering-Theory & Practice, 2014, 34 (9): 2417-2431.)